Second language Korean UD treebank v1.2:
Focus on data augmentation and annotation scheme refinement

Hakyung Sung[†] Gyu-Ho Shin[††]

[†]Linguistics, University of Oregon
[††]Linguistics, University of Illinois Chicago

RESOURCEFUL-2025

Research background

- **UD treebanks** have gained traction in **learner corpus research**: English (Berzak et al., 2016; Huang et al., 2018; Kyle et al., 2022; Lyashevskaya & Panteleeva, 2017), Chinese (Lee et al., 2017), Italian (Di Nuovo et al., 2019, 2022), Russian (Rozovskaya, 2024), and Swedish (Masciolini, 2023; Masciolini et al., 2023, 2024)
- Recent studies have developed **second language (L2) Korean** treebanks incorporating language-specific morphemes and dependency tags (Sung & Shin, 2023a, 2023b, 2024)

Research gaps

- Annotation needs iterative updates to balance cross-linguistic standardization with preserving language-specific features (De Marneffe et al., 2021; Manning, 2011)
  - Focused on language-specific features (Sung & Shin, 2024)
  - Towards **cross-linguistic standardization**

Key contributions

- **1. Data augmentation**: Augmented L2-Korean UD treebank (v1.1: 7,530 sentences) by adding 5,447 (*5,454) manually annotated sentences
- **2. Guideline revision**: Extensively revised dependency annotation guidelines to align with the UD framework, with minor adjustments for Korean-specific linguistic properties
- **3. Korean model fine-tuning**: Fine-tuned and evaluated three Korean language models in both in-domain and out-of-domain L2-Korean contexts using the updated treebank

Dataset	Metric	Baseline	Stanza	spaCy	Trankit
L2K-UD-test (in-domain)	XPOS	82.44	89.72	83.15	**91.81**
	LEMMA	89.61	**95.64**	87.97	88.84
	UAS	76.72	85.53	82.21	**92.28**
	LAS	60.69	80.36	75.21	**89.13**
KoLLA (out-of-domain)	XPOS	77.79	81.87	71.21	**84.51**
	LEMMA	88.03	**91.01**	79.64	86.90
	UAS	72.30	81.17	74.48	**88.93**
	LAS	58.53	75.14	63.56	**85.45**

Future directions

- Goal: Develop compact, high-quality L2-Korean treebank to fine-tune models and assess L2-Korean morphosyntactic analysis
  - Combine L2-Korean data from various genres and learner backgrounds
  - Introduce more language-specific features
  - **Currently:** Refine the alignment between universal **UPOS** and language-specific **XPOS** tags via expert revisions + L2-Korean data from **different domain**

More Information

Repository:
https://github.com/NLPxL2Korean/UD-KSL
(includes updated treebank, annotation guidelines, models, etc.)

UD-dev-branch:
https://github.com/UniversalDependencies/UD_Korean-KSL/tree/dev

Contact: hsung@uoregon.edu

References I

Berzak, Y., Kenney, J., Spadine, C., Wang, J. X., Lam, L., Mori, K. S., Garza, S., & Katz, B. (2016).Universal dependencies for learner english. *arXiv preprint arXiv:1605.04278*.

De Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021).Universal dependencies. *Computational linguistics*, *47*(2), 255–308.

Di Nuovo, E., Bosco, C., Mazzei, A., Sanguinetti, M., et al. (2019).Towards an italian learner treebank in universal dependencies. *CEUR workshop proceedings*, *2481*, 1–6.

Di Nuovo, E., Sanguinetti, M., Mazzei, A., Corino, E., & Bosco, C. (2022).Valico-ud: Treebanking an italian learner corpus in universal dependencies. *IJCoL. Italian Journal of Computational Linguistics*, *8*(8-1).

Huang, Y., Murakami, A., Alexopoulou, T., & Korhonen, A. (2018).Dependency parsing of learner english. *International Journal of Corpus Linguistics*, *23*(1), 28–54.

Kyle, K., Eguchi, M., Miller, A., & Sither, T. (2022).A dependency treebank of spoken second language english. *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, 39–45.

Lee, J. S., Leung, H., & Li, K. (2017).Towards universal dependencies for learner chinese. *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, 67–71.

References II

Lyashevskaya, O., & Panteleeva, I. (2017).Realec learner treebank: Annotation principles and evaluation of automatic parsing. *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, 80–87.

Manning, C. D. (2011).Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? *International conference on intelligent text processing and computational linguistics*, 171–189.

Masciolini, A. (2023).A query engine for l1-l2 parallel dependency treebanks. *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, 574–587.

Masciolini, A., Francis, E., & Szawerna, M. I. (2024).Synthetic-error augmented parsing of swedish as a second language: Experiments with word order. *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD)@ LREC-COLING 2024*, 43–49.

Masciolini, A., Volodina, E., & Dannlls, D. (2023).Towards automatically extracting morphosyntactical error patterns from l1-l2 parallel dependency treebanks. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, 585–597.

Rozovskaya, A. (2024).Universal dependencies for learner russian. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 17112–17119.

References III

Sung, H., & Shin, G.-H. (2023a). Diversifying language models for lesser-studied languages and language-usage contexts: A case of second language korean. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 11461–11473.

Sung, H., & Shin, G.-H. (2023b). Towards l2-friendly pipelines for learner corpora: A case of written production by l2-korean learners. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, 72–82.

Sung, H., & Shin, G.-H. (2024). Constructing a dependency treebank for second language learners of korean. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 3747–3758.